AD-A047 817 MARYLAND UNIV COLLEGE PARK COMPUTER SCIENCE CENTER F/G 6/4
UNDERSTANDING STICK FIGURES.(U)
NOV 77 M HERMAN
UNCLASSIFIED TR-603 N00014-76-C-0477
NL

1 OF 1
AD A047817

END
DATE FILMED
1-78
DDC

(12)
B.S.

COMPUTER SCIENCE TECHNICAL REPORT SERIES

UNIVERSITY OF MARYLAND
COLLEGE PARK, MARYLAND
20742

TR-603 November 1977
N00014-76C-0477

UNDERSTANDING STICK FIGURES

Martin Herman
Computer Science Center
University of Maryland
College Park, MD 20742

ABSTRACT

This paper describes a framework for a computer model of stick figure understanding in which semantic inferences are made from body postures. Stick figures of humans are shown to provide a rich environment in which to develop and test techniques and methods of visual understanding. The computer system which will implement the theory is called SKELETUN. The notion of a hierarchical organization of recognition packets running in parallel is introduced. This paradigm is applied to an example. In addition, it is shown how SKELETUN relies on interaction between model-driven and data-driven modes.

The support of the Information Systems Program, Office of Naval Research, under Contract N00014-76C-0477, is gratefully acknowledged, as is the guidance of Profs. C. J. Rieger and A. Rosenfeld.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
(6) UNDERSTANDING STICK FIGURES.

5. TYPE OF REPORT & PERIOD COVERED
(7) TECHNICAL rept.

6. PERFORMING ORG. REPORT NUMBER
(14) TR-603

7. AUTHOR(s)
(10) Martin/Herman

8. CONTRACT OR GRANT NUMBER(s)
(15) N00014-76C-0477

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Computer Science Center
University of Maryland
College Park, MD 20742

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
Information Systems Branch
Office of Naval Research
Washington, DC 20305

(11) 12. REPORT DATE
November 1977

13. NUMBER OF PAGES

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
(12) 44p.

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Image understanding
Stick figures
Visual semantics

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper describes a framework for a computer model of stick figure understanding in which semantic inferences are made from body postures. Stick figures of humans are shown to provide a rich environment in which to develop and test techniques and methods of visual understanding. The computer system which will implement the theory is called SKELETUN. The notion of a hierarchical organization of recognition packets running in parallel is introduced. This paradigm is applied to an example. In addition, it is shown how SKELETUN relies on

over

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

403 018

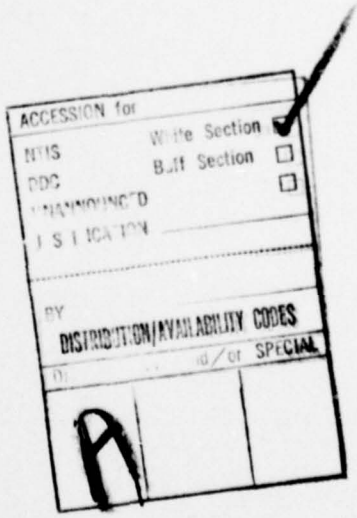
LB

ABSTRACT (contd.)

interaction between model-driven and data-driven modes.

Understanding Stick Figures

Table of Contents



1. Introduction

The goal of many researchers in artificial intelligence is to produce systems which exhibit capabilities that are normally attributed to human beings. If these capabilities are integrated into a single system, we would hope that the system would be able to understand and intelligently interact with a complex environment. The complex environment would consist not only of objects but also of people. An intelligent system will therefore have to be able to look at people and understand what they are doing, i.e., understand their actions and gestures. This includes perceiving and understanding the actions and gestures of isolated individuals as well as understanding the actions and gestures of groups of individuals interacting with each other. Examples of actions and gestures are running, throwing, pointing, fighting, and playing football.

What is involved in perceiving and understanding actions and gestures of people? First, a photograph or television picture of the people must be digitized and stored in memory. Then, features such as shapes of body parts, connected regions, and joint positions must be extracted. Next, these features must be combined in order to obtain a labeling of the body regions, e.g., upper arm, lower arm, torso, etc.

Once the body regions have been labeled, there are three main types of understanding which can occur. The first type involves understanding the actions and gestures of isolated individuals, assuming they are not interacting with other people. Local body gestures and body postures of the individual are used to infer what he is doing.

The second type involves understanding the interaction of two or more people. For example, if a photograph consists of one person with his hands stretched straight up and a second person holding a gun which is pointed at the first person, then our understanding of what the first person is doing (i.e., arms

stretched up) is increased if we take into account the action of the second person; that is, we hypothesize that a robbery may be occurring.

The third type of understanding involves taking a sequence of photographs of people and trying to understand the meaning of the whole sequence. This sequence may be either snapshots taken very rapidly one after another (e.g., 1/24 second apart), or scenes such as those in a comic strip which convey a story.

Implementation of the three types of understanding would be very difficult, since the technologies available for this purpose in the areas of computer vision and artificial intelligence are in their infancies. I have therefore decided to concentrate my research on understanding the semantics of body postures, body gestures, and actions, while avoiding the problems of low-level visual processing involved in extraction of syntactic features from pictures and labeling body parts in the picture. One way of avoiding these problems while still maintaining a rich domain of human gestures and actions is to limit the domain to stick figures of humans. This eliminates not only the problems of extracting, identifying, and labeling regions of the human body, but also the problem of occlusion of one part of the body by another.

My research will involve developing a theory and system which will recognize the actions and gestures of stick figures engaged in activities such as weeping, fighting, pointing, scratching, etc.

2. Motivation for the Research

Section 1 points out the desire of the artificial intelligence community to develop systems which can understand and interact with a complex environment. It is therefore important to develop techniques of understanding the gestures and actions of people, since people are certainly part of many environments. What are some of the specific applications, however, of this research? What purpose would the research serve?

One application lies in the field of robotics. A mobile robot, such as a robot house-servant, a robot baby-sitter, or a robot factory worker working among people, may need to have a relatively sophisticated capability of interacting with people. My research will develop techniques and methods which may help in providing this capability at the visual level.

Another purpose of this research is to develop computational techniques of understanding in general. Understanding systems are important in the natural language and problem solving domains as well as in the visual domain. Many of the issues involved in building understanding systems are common to all of these domains. I feel that the stick figure world offers a rich environment in which to test and develop computational techniques and methods of understanding. In this regard, my intention is not to develop an ad-hoc system which can understand only a limited set of stick figures. Rather, I intend to develop ideas and techniques which are general enough to be applicable to other domains of understanding. Examples of these domains might be scenes without people, natural language stories, and abstract concepts in the sciences and arts.

A final purpose of this research, and one that I feel is rather important, is to develop models of the cognitive processes used by the human mind in perceiving and understanding other humans in the visual world. Certainly, one can argue that any

system constructed to understand stick figures may have absolutely nothing to do with the way humans understand stick figures. Furthermore, it is extremely difficult, if not impossible, to use today's technology to test the validity of any model of human understanding. How then can anyone claim that a program running in a computer has anything to do with what goes on in the human mind?

The nature of cognitive processes is very poorly understood by psychologists. Since various methods of understanding these processes have yielded only modest results in the past, I feel that viewing the human mind as an information processing system is a valid approach which cannot be easily dismissed. If one accepts the possibility that this approach might be valid, then one can begin to comprehend why a computational system which understands stick figures might provide a means of investigating human understanding of stick figures. Certainly, I would not claim that any system which I construct works exactly the way the human mind works. What I do claim, however, is that the construction of such systems at least forces us to consider the details of what might be involved in cognitive processes. Because of this, the construction of such systems provides insight into the workings of the human mind which we might not have been able to obtain in any other way. The theory of understanding stick figures which I will develop will probably relate only marginally to the way humans understand stick figures. Hopefully, however, it will provide insight into the way the human mind accomplishes this understanding as well as providing a groundwork upon which better theories can be developed.

3. Related Work

The following work has been done in extracting information about human figures from pictures.

Speckert [S1] has looked at a sequence of pictures of a man walking as seen from a side view. His program extracts all the body joint positions (e.g., shoulder, elbow, hip, ankle, etc.) and labels them. This work complements my research, since it will aid in devising stick figure representations of humans.

Adler [A1] has built a system which can identify and give elementary descriptions of characters in Peanuts cartoons. His recognition process is concerned mainly with describing shapes of irregular, curved objects and finding hidden contours. Again, this work complements my research since it is concerned more with extracting and labeling body parts than with making semantic inferences.

Tsuji, Morizono, and Kuroda [TMK1] have attempted to analyze and understand a simple cartoon film. The input to their system consists of digitized versions of each frame in the film. The system is able to answer elementary questions about the action in the film. This work attempts to cover low-level visual processing of the film plus higher level inferences. Because of this, the system they have built is elementary in that it can analyze only one simple film. However, this is one of the first attempts to analyze and understand a dynamic world.

Some preliminary work in making semantic inferences in a visual domain has been done by Boose and Rieger [BR1]. They are concerned with extracting information from the facial expressions of characters in a children's story book. Facial features are used to infer qualities such as anger, happiness, and frustration. My research may be able to use some of the techniques and mechanisms developed by Boose and Rieger, although I am concerned primarily with expressions conveyed by the limbs of the body rather than expressions of the face.

Smoliar and Weber [SW1] are implementing a system which will produce animation of human movement on a graphics display. Their internal representation of human movement is in the format of Labanotation, a notation for recording human movement in dance. I will not be using an internal representation of stick figures based on Labanotation because I believe it is of no theoretical interest in the context of general human intelligence. Labanotation incorporates a detailed, symbolic, 3-dimensional, time-dependent representation which gives the position of every part of the body. It would be very difficult, perhaps impossible, to map an arbitrary 2-dimensional drawing of a stick figure into this representation, since the line segment lengths and angles are not accurately measured when drawn. I believe that a different approach (to be explained later) will prove more feasible and provide a better theoretical basis.

4. Research Proposal

My research will involve the construction and implementation of a theory of stick figure understanding. The computer implementation of this theory will be referred to as SKELETUN (SKELETon UNderstanding).

4.1. Syntax Component

A stick figure will consist of the following body parts: a circular head, two upper arms, two lower arms, two hands, two upper legs, two lower legs, two feet, and a torso consisting of three parts. Examples of stick figures are shown in Displays 1-10. Each part of the stick figure, except the head, will be a single straight line segment. The points at which the ends of the line segments meet will be the joints of the stick figure. Since the torso consists of three parts, it contains two joints.

The input to SKELETUN will consist of the following: (1) the coordinates of the end points of each line segment of the stick figure; (2) the points to which each point is connected by a line segment; (3) the coordinates of the center of the circle representing the head; and (4) the radius of this circle. This input will be hand-encoded by a human. Appendix A contains the LISP hand-encoded input for the figure in Display 1. After SKELETUN receives the input, it will label all body parts of the stick figure (i.e., head, upper arm, lower arm, etc.) and calculate the projected, 2-dimensional angles of all the joints (i.e., projected elbow joint angle, projected shoulder-upper arm joint angle, etc.). This initial preprocessing will provide input to the understanding component of SKELETUN.

4.2. Understanding Component

The understanding component will utilize the notion of frames [M1]. A frame is a data structure which represents a stereotype. It contains constraints which must be satisfied if an object (or circumstance) is to fall within the stereotype. Once an object can be seen to "fit" into a certain frame, strong predictions about properties of the object can be made without observing these properties directly, since the frame will suggest certain properties which are typical of its instantiations. The frame provides a general means for recognition. An additional property of frames is the following. The frames in a system are linked together in such a way as to form an information retrieval network. If a frame is invoked for recognizing a certain object, but the object fails some of its tests, then the network has the capability of suggesting other frames which may be invoked. Examples of frames in SKELETUN are SADNESS, RUBBING-FACE, SALUTING, LOOKING, SHOWING-OFF-MUSCLES, ARCHED-TORSO, SIDE-VIEW, TOUCHING, and others to be presented later. (The term "frame" will soon be abandoned in favor of a more specific term.)

4.3. Example

An example of the application of frames to the understanding of stick figures is the following. Suppose we input the drawing of Display 1 to SKELETUN. The initial preprocessing will label all body parts and determine projected angles of all the joints. Let us view the output of the preprocessor as data. SKELETUN will begin in a data-driven mode. At this point, data will be used to invoke frames. Examples of the kind of data which may be used for this purpose are: positions of the feet and hands, positions of the elbows and knees, shape of the torso, and others. The position of the feet in Display 1 will indicate that a side view of the figure is being observed, and thus a SIDE-VIEW

frame will be invoked. This frame will determine that the real angles at which the elbows are bent (i.e., in 3-dimensional space) are approximately the same as the elbow angles seen in the 2-dimensional drawing. Next, the positions of the hands inside the circle representing the head will invoke the TOUCHING frame. This frame contains heuristics to verify the occurrence of any touching relationships, especially touching relationships involving the hands and feet. Thus SKELETUN again becomes frame-driven, or model-driven. (The data- and model-driven concepts described here are similar to the concepts of active and passive knowledge described by Freuder [F2].) The TOUCHING frame has inherited the information from the SIDE-VIEW frame that the figure is facing the left and that its face is thus probably also pointing to the left. Since the hands are inside the left part of the circle representing the head, the TOUCHING frame concludes that the stick figure's hands are touching the face.

The information that the hands are touching the face, with the elbows in the given position, will invoke a number of candidate frames, including the SADNESS frame and the RUBBING-FACE frame. Suppose the SADNESS frame is given control initially. This frame will have information indicating which gestures are appropriate when the figure is sad. At this point SKELETUN will again become model-driven. The SADNESS frame will contain heuristics such as the following: "One of the things I should look for is a bent back or arched torso. I will therefore invoke the ARCHED-TORSO frame and see if it can find an arched torso." The ARCHED-TORSO frame will now gain control. It will have access to all data and results known by the SADNESS frame. The ARCHED-TORSO frame, knowing that the figure is being observed in a side view, will easily find the arched torso in the figure. When this feature is found, the ARCHED-TORSO frame will report success back to the SADNESS frame. If this information is enough to verify the SADNESS frame, then SKELETUN will "understand" this figure as being in a state of "sadness."

Suppose that SKELETUN is trying to understand Display 2. The positions of the feet, arms, and hands in this drawing are

similar to those in Display 1. SKELETUN would therefore again invoke the SADNESS frame. However, when the frame tries to verify that the figure has an arched back, the ARCHED-TORSO frame will instead report that the back is straight. Heuristics in the control component of SKELETUN (i.e., the information retrieval network) will indicate that if this occurs, control should be passed to the RUBBING-FACE frame. This frame will inherit all the data obtained by the SADNESS frame so that it will not have to duplicate any work. The frame will then attempt to determine whether or not the current stick figure "fits in." If this attempt is successful, SKELETUN will "understand" this stick figure as being in a state of "rubbing its face." Note that SKELETUN first tried SADNESS rather than RUBBING-FACE. The figures in both Displays 1 and 2 are rubbing their faces. However, saying that the figure in Display 1 is rubbing its face rather than being sad does not capture the most significant concept to be interpreted. Similarly, in Display 2, SKELETUN would have preferred to choose SADNESS. However, since SADNESS is inappropriate, RUBBING-FACE becomes the most significant concept.

This example of the use of frames points out one of its weak points, namely, the serial passing of control from one frame to another. The system would be more elegant and more general if it did not need a sequencing component to determine which frame gains control next. This could be implemented by conceptually making the system parallel rather than serial in its operation. The next section will elaborate upon this point.

5. Organization of the System

As explained above, the notion of frames will be heavily utilized by SKELETUN. So as not to lock myself into the concept of a "frame" as defined by Minsky, I will instead call it a "recognition packet" to emphasize its main function: recognition of semantically meaningful gestures and actions. The recognition packets as described here have some similarities to the packets of facts and demons described by Fahlman [F1].

5.1. Recognition Packets

SKELETUN will incorporate a hierarchy of recognition packets. The first version of SKELETUN (which will understand isolated stick figures) will consist of primary and secondary recognition packets. Recognition by the primary packets will constitute the main method of "understanding" by SKELETUN. These packets will consist of representations of semantic concepts. Thus, if a stick figure drawing can be categorized as showing SADNESS, or RUBBING-FACE, we will say that the system has achieved a high level of understanding. (See Section 7 for more examples of primary recognition packets.)

The secondary recognition packets will constitute a more syntactic-based, lower level of understanding. These will include ARCHED-TORSO, CHEST-THRUST-OUT, LEGS-STRADDLED, SIDE-VIEW, FRONT-VIEW, and TOUCHING. The secondary packets are components of recognition which by themselves are not adequate concepts for "understanding" the stick figures. However, the primary recognition packets may invoke or obtain results from the secondary packets to aid in the recognition process.

5.2. Parallelism in the System

Conceptually, SKELETUN will incorporate much parallelism in its processing. Initial data, for example, need not invoke only a single recognition packet; they may invoke several packets "simultaneously." Each packet will independently try to verify itself. Those which are unsuccessful will simply "die out." The others will declare themselves successful. It is important, however, that all recognition packets be able to communicate with each other. This can be implemented by means of a global work space. This concept is somewhat similar to the "blackboard" used in the Hearsay system [LE1]. The work space is an area which is accessible to all of the recognition packets. They will place their results in the work space and obtain the results found by other packets from the work space. In the example in Section 4, the SIDE-VIEW, ARCHED-TORSO, and TOUCHING packets will pass their results to the SADNESS and RUBBING-FACE packets via the work space. If one recognition packet needs to invoke another one, this can be done implicitly by placing a message in the work space.

The concept of the work space allows all recognition packets to be independent modules. Thus, if one wants to add a new packet to the system, he need not know about all of the other packets already in the system. Furthermore, the independence of recognition packets implies that the system can be run on a parallel machine, since there need be no higher level sequencing component to determine which packet must be invoked when. The notion of parallel recognition packets acting simultaneously is also theoretically pleasing since the human mind seems to work similarly.

5.3. Example of the Control Structure

Consider the example in Section 4 in light of this parallel

organization. Display 11 shows the hierarchy consisting of data at the lowest level, and secondary and primary recognition packets at the next two levels. Display 12 is what a trace of the program running on this example might look like. Either a recognition packet or data may be in control. When a recognition packet is in control, we say that the system is model-driven. As the packet attempts to verify itself, it will make assertions which are placed in the global work space. Recall that all assertions in the work space are accessible to all packets. When data are in control, we say that the system is data-driven. In Display 12, an assertion made by data indicates which data are invoking the models. The purpose of this display is to indicate the control used for this example. Display 11, on the other hand, indicates the organization of the system required for this example.

The numbers above the boxes in Display 11 indicate the order in which each recognition packet is invoked. Note that the three secondary recognition packets are all invoked by the data simultaneously. However, both ARCHED-TORSO and TOUCHING need to know which view of the figure is being observed. They therefore must wait for SIDE-VIEW to place this information in the work space before they can proceed. SADNESS and RUBBING-FACE are also invoked simultaneously, but only after TOUCHING and SIDE-VIEW have asserted the following: (1) the 2-dimensional angles of the elbows are almost the same as the 3-dimensional angles; and (2) the hands are touching the face. By this time, ARCHED-TORSO has already asserted that the torso of the figure is arched, and therefore SADNESS reports success while RUBBING-FACE reports failure.

5.4. Body Specialists

Instead of using only the output of the preprocessor as data (i.e., labels of body parts and projected angles of joints), I may try the following. Each body part of the stick figure, i.e.,

right foot, left foot, right hand, left hand, right upper arm, left upper arm, etc., can be represented by a specialist whose purpose is to produce a symbolic representation of the position of its body part in 3-dimensional space. In this manner, a foot may be represented as "pointing left," "pointing right," or "pointing forward" with respect to the viewer of the drawing. This would provide the system with two levels of "data." The first level would consist of the output of the preprocessor; the second level would consist of the symbolic representation of the positions of the body parts. In this way, the recognition packets may look at either level of data.

5.5. Notes on Implementation

SKELETUN will be implemented in LISP. It may incorporate "demons" in its implementation. Conceptually, a demon's purpose is to get "excited" when a particular set of events occurs. The demon "watches" an arena of events, and when the set of events for which it was created occurs, the demon executes a body of instructions. In SKELETUN, demons may be used for body part specialists or for invoking recognition packets. An interesting implementation of demons in the form of "spontaneous computations" has been done by Rieger [R2]. I may adapt this implementation for use in SKELETUN.

6. Stages in the Research

The proposed research may be divided into three stages. The first stage will involve constructing and implementing a theory of understanding isolated stick figures. This is the initial purpose of the research, and most of the ideas in this paper are directed toward that end.

The second stage of this research will involve adding the capability of understanding two or more stick figures which are interacting. Consider the example provided earlier, consisting of one figure standing with its arms straight up in the air and another figure holding a gun pointed towards the first figure. If one were to attempt to understand each figure individually, a description of the scene might be: "One figure is stretching and the other is pointing a gun." However, an understanding involving the two figures as interacting humans might be: "A robbery is occurring!" Indeed, the understanding of the scene changed when the total context was taken into account.

Other examples of high level semantic inferences that can be made about groups of stick figures are (1) "chorus line" instead of "a group of dancers;" (2) "track meet" instead of "a bunch of people running;" and (3) "soccer game" instead of "a bunch of people running."

The third stage of this research will involve investigating the understanding of comic strips which are purely pictorial and consist only of stick-figure characters. This would probably use many of the techniques used in natural language story comprehension [R1], since a pictorial comic strip tells a story. A picture-story understanding system will be able to infer all of the cause-effect relationships which occur from one picture frame of the comic strip to another. The system will display its understanding of the comic strip by outputing an English summary of the story line.

7. The Data Base

The data base consists of drawings of stick figures by many different people. Displays 1-10 show examples of drawings in the data base. The first version of SKELETUN will have ten primary recognition packets to be used for understanding. The names of these packets, plus examples of stick figures whose gestures can be categorized by these packets, are the following:

(1) weeping or sad - Display 1
(2) rubbing face - Display 2
(3) relaxed position - Display 3
(4) saluting - Display 4
(5) looking or seeking - Display 5
(6) pointing - Display 6
(7) showing off muscles - Display 7
(8) cupping ear to hear better - Display 8
(9) boxing stance - Display 9
(10) scratching own back - Display 10

Note that most of the recognition packets deal with static gestures rather than actions, particularly vigorous actions. I feel it is best to start with static gestures since they are easier to understand.

The examples shown in Displays 1-10 are only part of the whole data base presently available. This data base was recently obtained from an introductory drawing class at University of Maryland. A brief description of each of the ten concepts was given to the forty students in the class. Each student was then asked to draw stick figures which they thought would instantiate each concept.

Each drawing will be presented to SKELETUN. If the program chooses one of the ten recognition packets, then we can say that it has a good "understanding" of the drawing. The program may choose more than one packet, in which case it will be saying:

"I'm not really sure what's going on in the drawing, but it looks as if the stick figure is doing something which falls somewhere within these packets." If it chooses no packet, this means that SKELETUN cannot "understand" the drawing based on what is presently in its memory. This is acceptable since people often have the same problem, i.e., people quite often cannot really understand what a stick figure is doing, since it might be doing something which cannot be discerned, or it may be doing nothing in particular. When this case arises, SKELETUN will try to provide a concise description of the figure by using the secondary recognition packets which were instantiated. For example, SKELETUN might output: "This figure has its left leg up in the air, its chest thrust out, and its hands clasped on top of its head." This kind of description shows a partial understanding which is based more on the syntax of the stick figure than on the semantics. In the second version of SKELETUN, where drawings will involve the interactions of two or more stick figures, partial descriptions of each stick figure, when combined under a higher level recognition packet, might result in a fuller understanding of the scene.

8. Testing Methodology

The theory of understanding which I will construct, plus its implementation through SKELETUN, will need to be tested to determine its adequacy. In order to test the first version of SKELETUN, I will obtain a new data base of drawings from both amateurs and students of art. In addition, the data base will contain drawings of stick figures derived from cartoons and photographs. The drawings will be done by people unfamiliar with the program. The same set of drawings will then be presented to a group of subjects who also are unfamiliar with the program and who have not seen the drawings before. The subjects will be asked to concisely describe what each stick figure is doing. These descriptions will be compared to the recognition packets or brief descriptions chosen by SKELETUN. For each stick figure in the data base, if the subjects choose a description which is similar to one of the concepts represented by a recognition packet in SKELETUN, then we would expect SKELETUN to choose that same recognition packet.

9. Conclusions

The domain of stick figures is rich enough to provide a good environment in which to develop and test techniques and methods of visual perception and understanding. Furthermore, the domain of stick figures is interesting in its own right; that is, a theory and computer system for stick figure understanding can be applied to practical engineering problems of robot interaction with humans. Finally, since the domain of stick figures has been relatively unexplored, research in this area is overdue.

Acknowledgment

I would like to thank Prof. Cynthia Bickley for allowing me to use her drawing class to obtain the data base of stick figures.

10. References

[A1] Adler,M., Computer Interpretation of Peanuts Cartoons, 5th IJCAI, Aug. 1977.

[BR1] Boose,J. and Rieger,C., The Windows to the Soul: Character Behavior from Visual Inferencing in a Children's Story, University of Maryland TR 579, July 1977.

[F1] Fahlman,S., A Hypothesis-Frame System for Recognition Problems, MIT AI Lab Working Paper 57, Dec. 1973.

[F2] Freuder,E., A Computer System for Visual Recognition using Active Knowledge, 5th IJCAI, Aug. 1977.

[LE1] Lesser, V. R., and Erman, L. D., A Retrospective View of the Hearsay-II Architecture, 5th IJCAI, Aug. 1977.

[M1] Minsky,M., A Framework for Representing Knowledge, in The Psychology of Computer Vision, P.H.Winston (ed.), McGraw-Hill, 1975.

[R1] Rieger,C., GRIND-1: First Report on the Magic Grinder Story Comprehension Project, University of Maryland TR 588, Oct. 1977.

[R2] Rieger, C., Spontaneous Computation in Cognitive Models, University of Maryland TR 459, July 1976.

[S1] Speckert,G., Knowledge Driven Recognition of the Human Body, MIT Working Paper 118, Jan. 1976.

[SW1] Smoliar,S.W., and Weber,L., Dance Notation and the Computer, submitted to Artificial Intelligence.

[TMK1] Tsuji,S., Morizono,A., and Kuroda,S., Understanding a Simple Cartoon Film by a Computer Vision System, 5th

IJCAI, Aug. 1977.

11. Appendix A

This appendix shows the hand-encoded input to SKELETUN of the stick figure in Display 1. The coding is in LISP. The stick figure contains 17 end points of the line segments, plus one point for the center of the circle representing the head. The end points are labelled from 1 to 17. The x and y coordinates of each point are determined based on coordinate axes whose origin and orientation is arbitrarily chosen. This procedure is performed by a human rather than by a program. Note that each end point shares line segments with at least one other point.

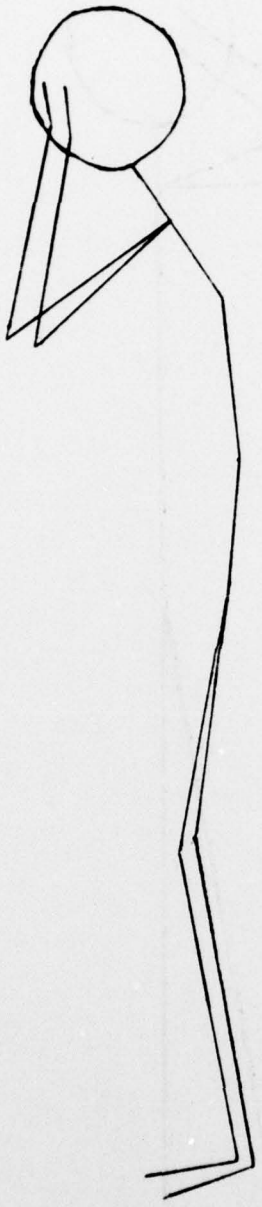
The circle is represented as follows:
(<x coord. of ctr.> <y coord. of ctr.> <radius>)

Each end point is represented as follows:
(< point no.> (<x coord.> <y coord.>) <a list
of the other points it is connected to>)

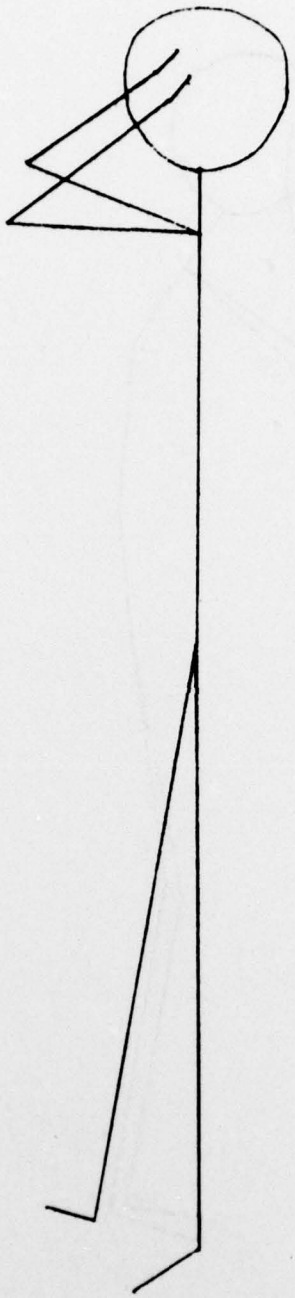
LISP Input

```
(( 9.4   65.5   4.7 )        <---- circle
 ( 1   ( 8.8   2.2 )   ( 3 ))
 ( 2   ( 7.4   3.5 )   ( 4 ))
 ( 3   ( 14.0   3.7 )   ( 1   6 ))
 ( 4   ( 13.0   3.9 )   ( 2   5 ))
 ( 5   ( 10.9   21.4 )   ( 4   7 ))
 ( 6   ( 11.5   22.5 )   ( 3   7 ))
 ( 7   ( 14.3   36.0 )   ( 5   6   8 ))
 ( 8   ( 15.4   43.8 )   ( 7   11 ))
 ( 9   ( 2.7   51.5 )   ( 12   15 ))
 ( 10   ( 4.3   51.0 )   ( 12   16 ))
```

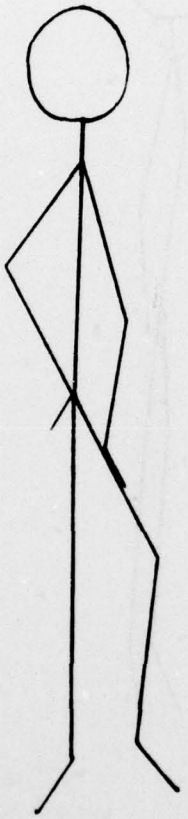
```
( 11  ( 15.1  52.0 )  ( 8  12 ))
( 12  ( 12.4  5c.0 )  ( 11  10  9  17 ))
( 13  ( 6.9  65.5 )  ( 15 ))
( 14  ( 7.7  c5.7 )  ( 16 ))
( 15  ( 6.6  63.2 )  ( 9  13 ))
( 16  ( 7.8  62.1 )  ( 10  14 ))
( 17  ( 11.0  61.2 )  ( 12 )))
```



DISPLAY 1



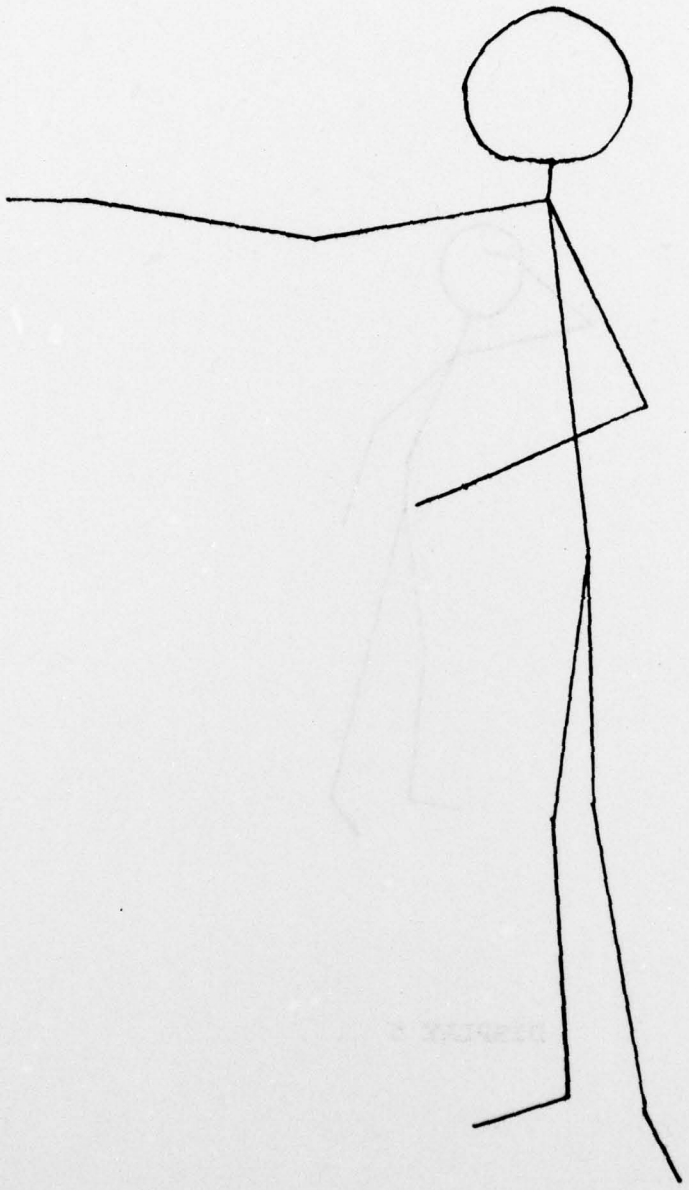
DISPLAY 2



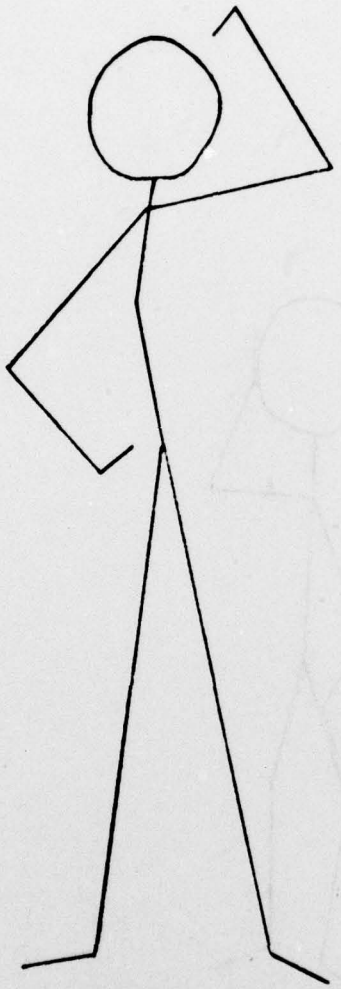
DISPLAY 3

DISPLAY 4

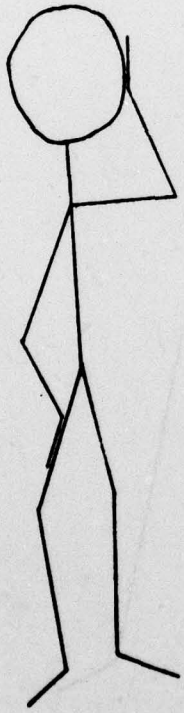
DISPLAY 5



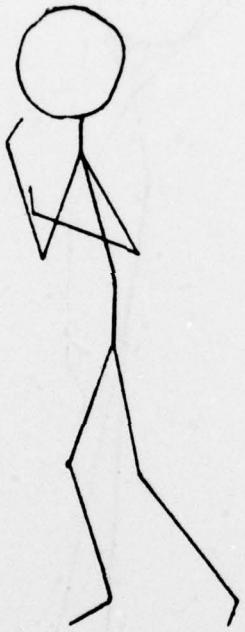
DISPLAY 6



DISPLAY 7

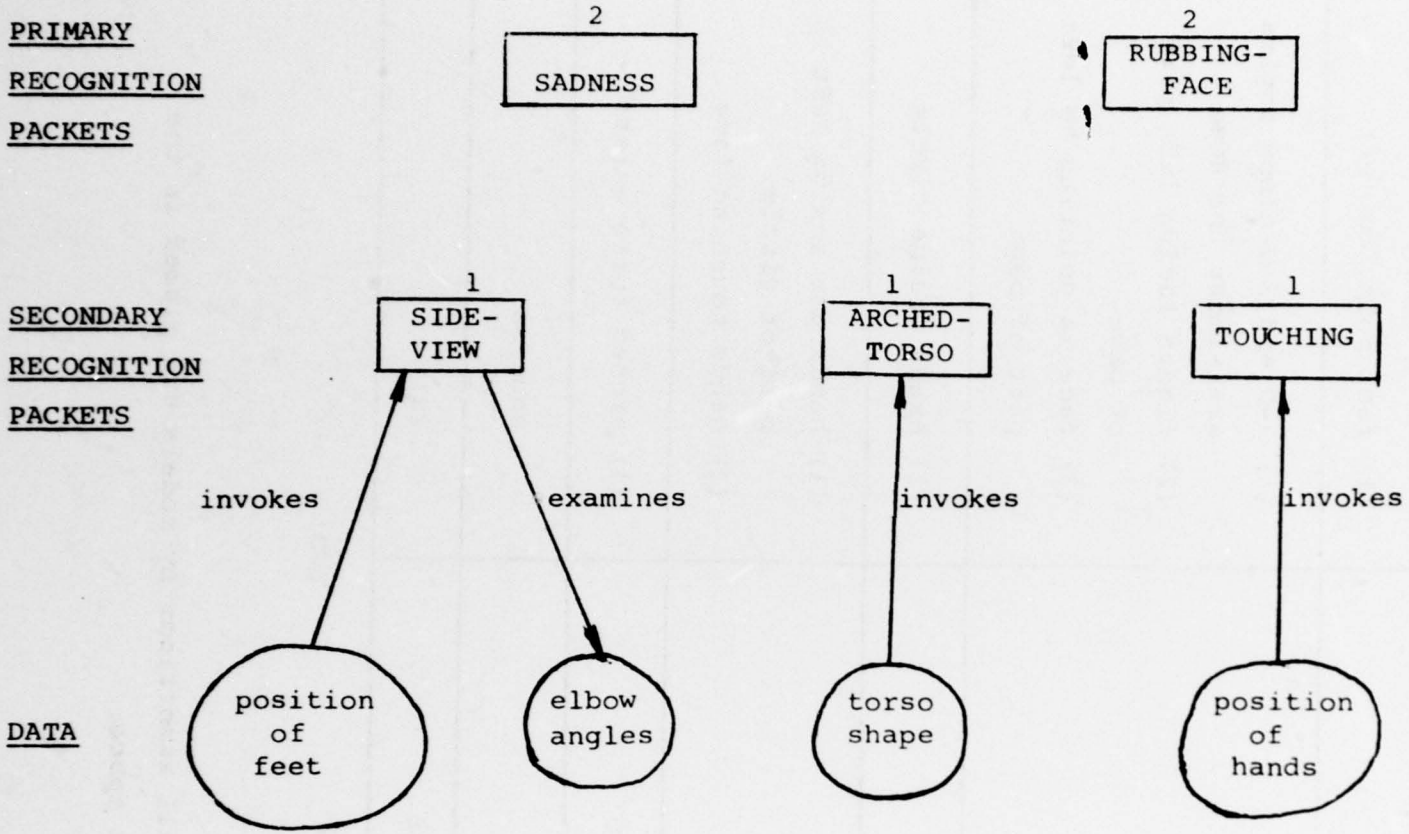


DISPLAY 8



DISPLAY 9

DISPLAY 10


PRIMARY
RECOGNITION
PACKETS
2
SADNESS
2
RUBBING-
FACE
SECONDARY
RECOGNITION
PACKETS
1
SIDE-
VIEW
1
ARCHED-
TORSO
1
TOUCHING
invokes
examines
invokes
invokes
DATA
position
of
feet
elbow
angles
torso
shape
position
of
hands

DISPLAY 11

IN CONTROL	ASSERTIONS
Data	(1) feet pointing left
SIDE-VIEW	(1) 3-D and 2-D elbow angles are almost the same (2) figure facing left part of page (3) face is pointing to left part of page
Data	(1) hands inside circle
TOUCHING	(1) hands are inside left part of circle (2) hands touching face
ARCHED-TORSO	(1) arched torso exists
SADNESS	success
RUBBING-FACE	fail

END

Note that all assertions by models are placed in the global work space.

DISPLAY 12

Off of Naval Research
Branch Office, Boston
495 Summer St.
Boston, Mass. 02210

New York Area Office
715 Broadway-5th Floor
New York, N.Y. 10003

Mr. E. H. Gleissner
Naval Ship R+D Center
Computation and Math Department
Code 18
Bethesda, Maryland 20084

Capt. Grace M. Hopper
NAICOM/MIS Planning Branch
OP-916D
Off, Chf. of Naval Op.
Washington, D.C. 20350

Mr. Kin B. Thompson
Technical Director
Information Systems Div. OP-91T
Off., Chf. of Naval Op.
Washington, D.C. 20375

Naval Research Lab.
Technical Info. Division
Code 2627
Washington, D.C. 20375

Dr. A.L. Slafkosky
Scientific Advisor
Commandant, USMC
Code RD-1
Washington, D.C. 20380

National Security Agcy.
Attn: Dr. Maar
Fort Meade, Maryland 20755

Off. of Naval Research
Code 1021P
Arlington, Va. 22217

Asst. Chief for Tech.
ONR Dept. of Navy
Code 200
Arlington, Va. 22217

Off. of Naval Research
Information Sys. Program
Code 437
Arlington, Va. 22217

Off. of Naval Research
Code 455
Arlington, Va. 22217

Off. of Naval Research
Code 458
Arlington, Va. 22217

Defense Documentn. Cent.
Cameron Station
Alexandria, Va. 22314

Off. of Naval Research
Branch Office, Chicago
536 South Clark St.
Chicago, Ill. 60605

Off. of Naval Research
Branch Off., Pasadena
1030 East Green St.
Pasadena, Calif. 91106

Naval Electron. Lab. Ctr.
Adv. Software Tech. Div.
Code 5200
San Diego, Calif. 92152